Naveed Jhamat[1], Hafiz Amaad[2], Ghulam Mustafa[3, *], Naeem A. Nawaz[4], Muzaffar Bashir[5] and Muhammad Atif Makhdoom[6]

Entropic Distance-based Classification of Tourist Attractions Using K* Learner: An Instance-based Approach

Abstract

The classification of tourism attractions is becoming a more well admirable research trend. Various techniques have been taken on board to classify tourism attractions to provide information to the users based on their preferences. This study provides an in-depth insight into the entropic distance-based classification approach for the prediction of user attractions. In this study, a lazy classification technique, k-star, is implemented to predict the tourism places based on user ratings. The k-star algorithm is the nearest neighbor approach that discovers the nearest instances to the targeted instance. Unlike other nearest neighbor approaches, the k-star algorithm exploits entropic distance, which measures all the possible shortest paths to discover the nearest instances based on user ratings. Furthermore, the evaluation assessments are also carried out to justify the performance of the k-star algorithm.

Keywords: Tourism, online service, ratings, trip advisor, dataset, k-star algorithm

1. Introduction

In recent years, tourism attractions are becoming famous progressively. A large number of tourists prefer online services to book hotels and search attractions to visit rather than reserving whole tours. In this respect, various approaches toward information retrieval, which allow searching information about tourism trips as well as provide information about tourism attractions during the journey, are becoming ever more popular (Smirnov et al., 2013). Several web-based services are also facilitating customers for discovering tourism attractions based on their interests. Several online services have been developed using the different web and mobile technologies that provide easy and useful functionalities to classify extensive data of tourist attractions worldwide. Although, these services have to personify the tourist’s preferences before providing the list of tourist attractions. The main objective of the systems providing such services is to discover and identify which attractions are suitable for the user for attendance (Smirnov et al., 2013).

1.1 Motivation and Aim

In this study, a classification technique, lazy k-star, for tourism attractions is presented, which predicts tourism attractions based on the user ratings given on a particular tourism attribute such as museums, restaurants, parks, resorts, etc. This technique predicts interesting tourism places according to user preferences. In this regard, we intend to apply the lazy k-star algorithm on a real-world dataset, TripAdvisor, to classify different tourism instances based on provided ratings by different users. Moreover, we aim to provide a synopsis of our dataset and a demonstration of the k-star algorithm and an alternative algorithm closest to the k-star algorithm in terms of performance.

2. Related Work

In the work of Kashevnik et al. (2017), A multimodel approach to the implementation of context-aware recommender systems in the field of tourism knowledge support is provided. Recommendation approaches apply to personal knowledge and non-personalized, but a recommendation module must be constructed to adapt to particular circumstances. Similarly, Saputra et al. (2019) intended to demonstrate how the supporting details like the database layout, the data design, and the data representation in the tourism recommendation framework are gathered. The study outcome may be used for more analysis - specific on tourism recommendation method considering the change in weather or traffic situation and its impact on traveling or a trip. Khallouki et al. (2018) implemented a modern method for developing mobile tourism recommendation systems. IoT technologies are paired with a semantic web service to forecast the tourist's real-time context and provide the necessary services. According to Lee et al. (2017), it is easy to locate traveling knowledge on the internet by typing in keywords. Despite this being the primary goal of an approach, the consumers' input would be analyzed unacceptably. They used the Semantic Web and SPARQL to construct the Tourism Ontology and applied Fuseki to question the Ontology. Its recommender system provides details on attractive locations to go to and a friendly experience. Recommender scheme targets the tourism user, their accommodation, and the desirable locations. Similarly, Kularbphettong and Ngamkam (2014) proposed a hybrid recommender framework based on ontologies for suggesting heritage-tourism sites in Thailand to facilitate tourist users to search, make choices, and schedule their travel. The mobile application is designed upon a complex structure, and the mobile application is being improved continuously. Recommending locations is a vital task for smartphone devices since it requires the users' current location and the exploration of the most desirable places. They used an ontology, location-related facilities, and interactive search algorithm to suggest products based on tastes. Furthermore, they employed a graph to connect the tourist option with the specified trip by a distance function. Evaluations of their method show that their recommender system is able to suggest sites depending on the needs of the visitor and it is the right option for them. The work of Kavitha et al. (2017) collected twitter metadata, executed a Latent Dirichlet Allocation algorithm, and provided a collection of subject probabilities. A destination tree is constructed from India's tourist sites, the leaves reflecting a tourist site in India. Focused on the user node, the tourist tree is designed to suggest tourist destinations in India.

3. TripAdvisor Dataset

A real-world dataset known as TripAdvisor (Renjith, 2018) is used in this study to classify tourist attractions by implementing the k-star algorithm. The TripAdvisor dataset contains ratings of 980 users on different tourism attractions. This dataset has 981 instances/rows, including column headers, ten tourism attributes/columns of numerical type. Moreover, the dataset also contains a unique ID for each user of the nominal type. Categorically, the dataset contains 11 columns and 981 rows; the columns can be distinguished as both the numerical and nominal type where ten columns are numerical that comprise ratings of tourism attractions/attributes while only one column is nominal, which contains the unique ID of each user ranged from User1 – User980. The tourism attributes comprise ratings of 10 tourism attractions.

The dataset was downloaded from the online machine learning dataset repository (UCI, 2018) and contained several ratings against each tourism attraction (Table 1). Considering the tourism domain, the dataset was initially generated from different corresponding dimensions of user interests such as reviews, feedbacks, and ratings acquired from different social media channels (Renjith et al., 2018). Table 1 and Figure 1 provide the detail about the TripAdvisor dataset in different aspects.

Table 1 *A synopsis of TripAdvisor Dataset, Total Ratings = 9800*

Cat. No	Avg. Ratings on	Distinct Ratings	Min	Max	Mean	Std. Deviation
1	Art galleries	84	0.34	3.22	0.893	0.327
2	Dance clubs	73	**0**	3.64	1.353	0.478
3	Juice bars	**195**	0.13	3.62	1.013	**0.789**
4	Restaurants	91	0.15	3.44	**0.532**	0.28
5	Museums	87	0.06	3.3	0.94	0.437
6	Resorts	135	0.14	**3.76**	1.843	0.54
7	Parks/picnic spots	**6**	**3.16**	3.21	**3.181**	**0.008**
8	Beaches	68	2.42	3.39	2.835	0.138
9	Theatres	65	0.74	**3.17**	1.569	0.365

10	Religious institutions	56	2.14	3.66	2.799	0.321
Total	**N/A**	**860**	**9.28**	**34.41**	**16.958**	**3.683**

Table 1 illustrates information about the TripAdvisor dataset used in this work. The synopsis given in Table 1 illustrates that juice bars have highly distinct ratings while parks and picnic spots have the lowest tendency of distinctness. Similarly, restaurants have the lowest mean average of overall ratings, while parks and picnic spots have the highest mean value. On the other hand, parks and picnic spots have the lowest standard deviation, and juice bars have the highest standard deviation.

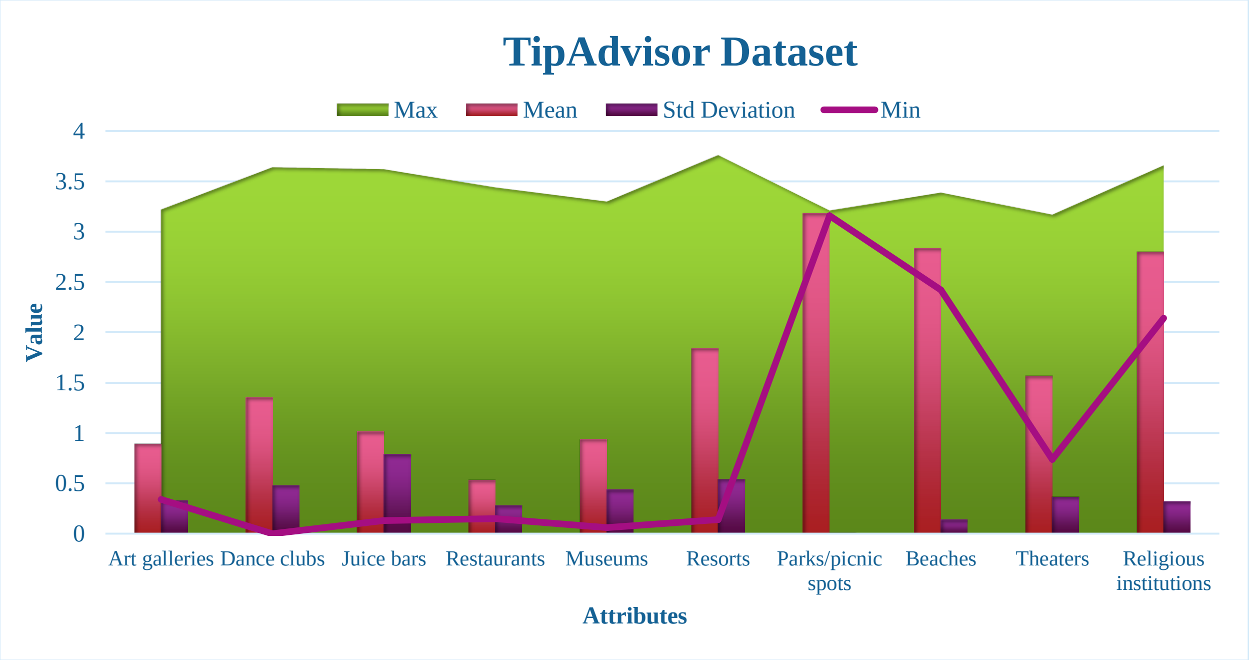

Figure 1 *Visualization of TripAdvisor Dataset*

Furthermore, for better understanding, the findings of Table 1 are portrayed in Figure 1 in terms of minimum and maximum ratings and mean and standard deviation against each tourism attraction.

4. K* Algorithm

In this work, we intend to implement the k-star algorithm, coined by Cleary in 1995 (Cleary and Trigg, 1995), on the TripAdvisor dataset for the classification of tourism attractions. This algorithm was chosen for the implementation on the TripAdvisor dataset in this study after a chain of experiments. In accordance with the nature of the dataset, to the best of our knowledge, among other classification algorithms available in Weka, the k-star algorithm was ascertained as the most suitable algorithm for the classification of tourism attractions using the TripAdvisor dataset. K-star algorithm is an instance-based learner/classifier where each targeted instance is compared with other existing instances utilizing a distance function, and the closest one is used to assign the targeted class (Witten et al., 2016). It is a kind of nearest neighbor classifiers. The main difference between this classification technique and other nearest neighbor classification techniques is the use of entropy distance function, which is measured by the complexity of converting an instance into another for the classification/prediction of most suitable and appropriate instances based on the similarities between the two users (Cleary and Trigg, 1995). All of this process is done through the probability of conversions/transformations in a random manner (Tejera Hernández, 2015). The similarity is calculated through similarity tools such as ratings and feedbacks of the users. The K-star classifier predicts the user's future interests by calculating the similarity between two users of the same mind. Since past few years, the k-star algorithm is widely used by various researchers for the classification/recommendation purposes in different application domains (Bahri et al., 2013; Jegadeeshwaran and Sugumaran, 2014; Madhusudana et al., 2016; Painuli et al., 2014; Satishkumar and Sugumaran, 2017).

4.1 Entropy Distance

The approach to compute the distance between two neighbors or instances is based on the information theory (Jaynes, 1957). The logic behind the distance can be defined as the complexity of converting one instance into another. On the other side, the complexity is calculated in two phases:

1) A finite set of conversions is defined, which draws one instance to another instance.
2) A conversion of one instance *x* to another instance *y* is an approach that starts at *x* and terminates at *y*.

Generally, the complexity can be defined as the size of the shortest string linking two instances together (Li and Vitanyi, 1997). This technique considers only a single conversion which should be the shortest one out of numerous possible conversions. The outcome is the distance measure which is very complicated to minor changes in the instance space and does not solve the smoothness problem appropriately. The k-star distance efforts to tackle this problem by accumulating all possible conversions between two instances through entropy distance.

In simple words, the distance can be described as the probability of random selection of conversions. In regard to the entropy distance, it is the probability that an instance will arrive employing a random walk away from the actual instance. Once the overall paths are gathered, the probability can be converted into different complexity units by implementing the algorithm. The accumulation technique of overall possible conversions was first successfully used in the r-theory of Yee and Allison (1993), in which they measured the distance between DNA patterns. Moreover, it was empirically witnessed that the use of all mutational conversions between two instances instead of the single shortest path provided a more robust and realistic measure of relatedness between two DNA patterns (Cleary and Trigg, 1995).

4.2 Advantages and Disadvantages of K*

Some advantages of the k-star algorithm obtained from the study of Cleary and Trigg (1995) are as follows:

4.2.1 Dealing with missing values

The k-star algorithm deals with the missing values in the instances very efficiently and more effectively. If a dataset instance that is going to be classified contains some missing values, then the instance attributes can simply be ignored, and the predictions are generated only for the remaining attributes. Furthermore, if the missing values in an instance, which is stored in the database, the way to deal with this is to assume the existing values under the attribute to be randomly drawn to cover this gap.

4.2.2 Symbolic Probabilities

Another advantage of this approach is dealing with both the continuous and symbolic attributes within the same framework.

4.2.3 Real Numbers

This algorithm can deal with real numbers (positive and negative integers) easily.

4.2.4 Combining Attributes

In this algorithm, the computation of the distance between two instances having more than one attribute is straightforward. The set of conversions on the joined attributes can be dealt with as the union of the conversions for the individual attributes, and the conversion string can then be modeled in sequential order by converting the first attribute, then the second, and so on until the last attribute is converted.

The disadvantages highlighted by Venkata Ramana (2011) of this algorithm include:

1. Laziness
2. Memory limitation

3. Sensitive in a local data structure

Furthermore, a step-by-step workflow is portrayed in the following flowchart of the k-star algorithm.

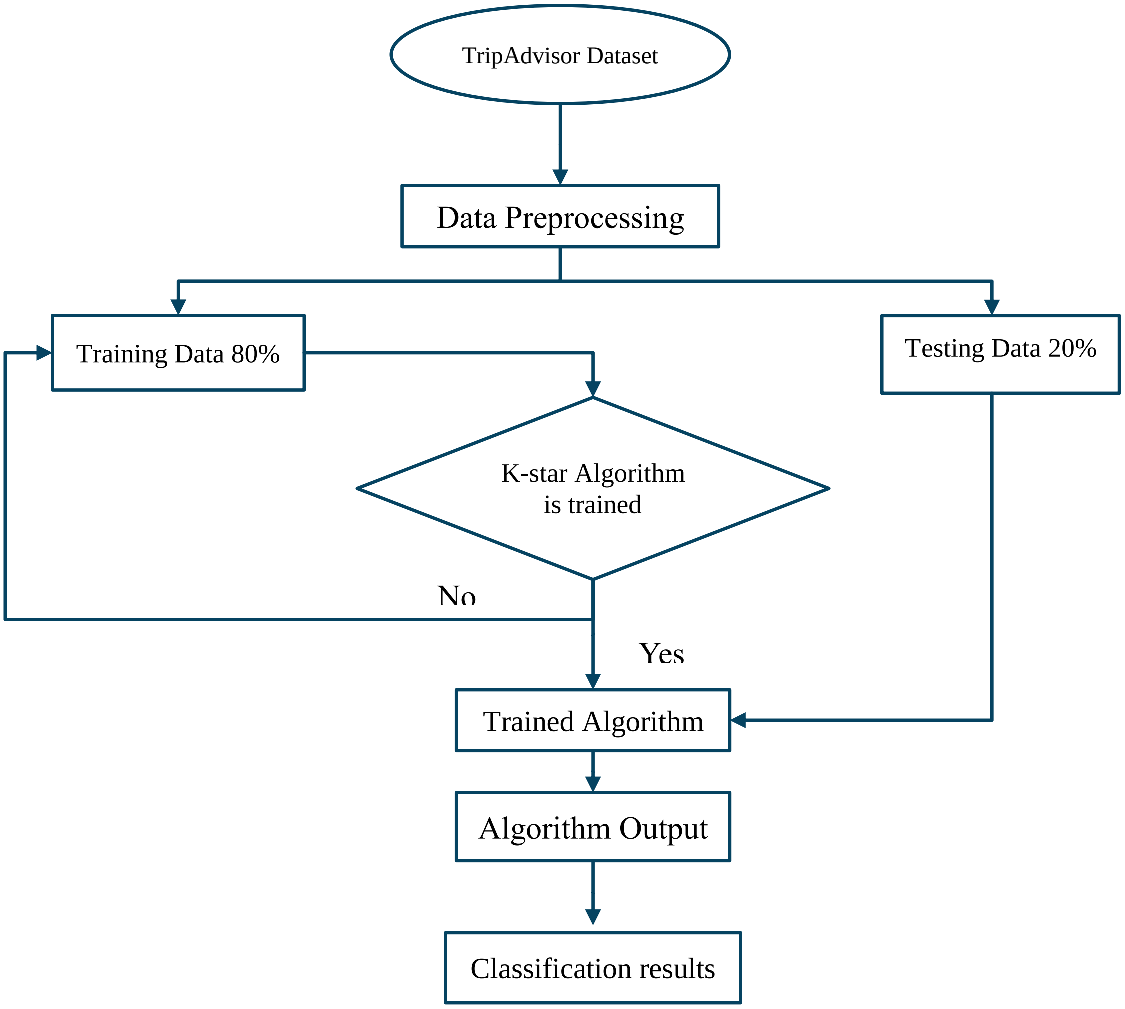

Figure 2 *Flowchart for the Classification of Tourism Attractions*

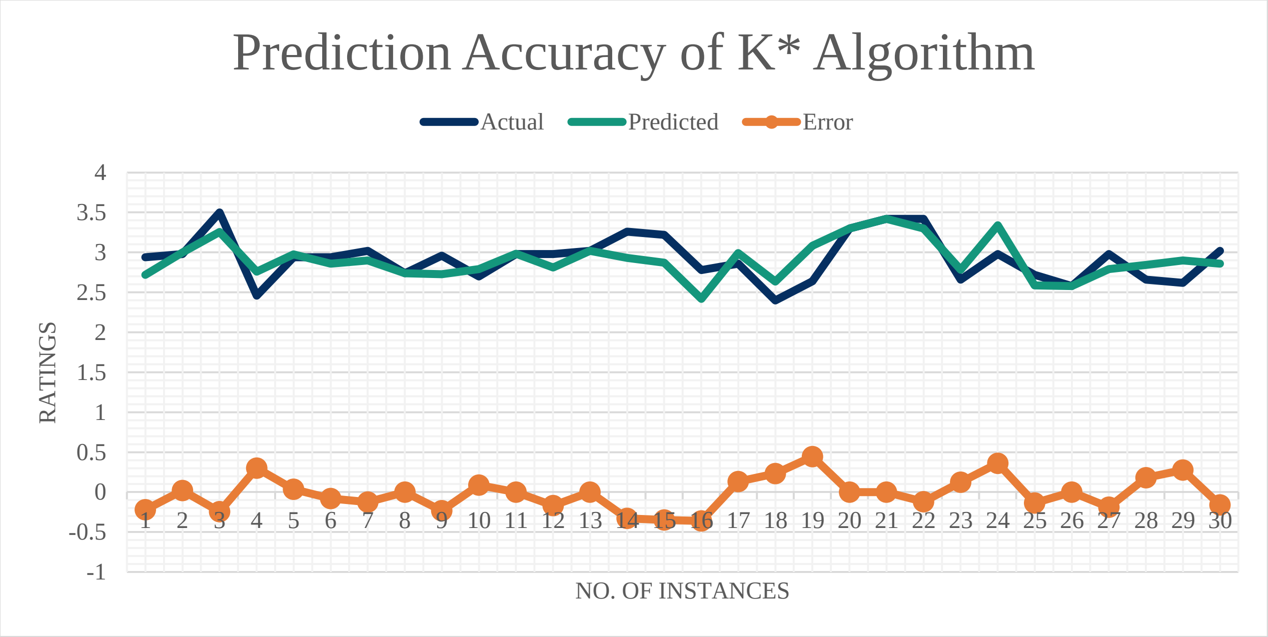

Figure 3 *Performance of K-Star Algorithm in terms of prediction*

Figure 3 portrays the predictions generated by the k-star algorithm against each particular instance. Both the actual and predicted ratings are illustrated in the figure. Additionally, an error score is also given for each instance in the figure.

Figure 3 demonstrates the closeness of predicted instances to the actual ones. Furthermore, section 4 demonstrates the experimental evaluations of this study. According to the experimental evaluations, an alternative approach to the k-star algorithm can be another lazy approach called Local Weighted Learning (LWL) which provides better results after the k-star algorithm to classify tourism attractions using the TripAdvisor dataset.

4.3 Pseudo Code of k-star Algorithms

```
K ← number of closest relevant instances
    For each instance $X$ in the testing set, do
        measure the entropic distance $ED$ between $X$ and every instance $Y$ in the training set
        the closeness ← the $k$ closest instances to the $X$ in the training set
        X.class ← SelectClass(closest instances)
    End for
```

5. Experimental Evaluation

In this section, the evaluation results of the k-star algorithm on the TripAdvisor dataset are demonstrated in comparison with other baseline approaches. The baseline algorithms include:

1) Linear Regression (LR)
2) Random Forest (RF)
3) Decision Stump (DS)

4) Locally Weighted Learning (LWL)

Furthermore, the evaluation metrics used to justify the performance of the k-star algorithm compared with other baseline approaches include:

1) Mean Absolute Error (MAE)
2) Root Mean Square Error (RMSE)
3) Relative Absolute Error (RAE)
4) Relative Squared Error (RSE)
5) Elapsed Training Time
6) Elapsed Testing Time

A synopsis of each baseline approach and evaluation metrics are provided in the next subsections.

5.1 Baseline Approaches for Evaluation

5.1.1 Linear Regression

Linear Regression is a primary and most broadly used algorithm for classification/predictive analysis. The common intention of linear regression is to examine two aspects:

1) Do the predictor instances provide good performance for predicting an outcome?
2) Which instances are the significant predictor of the outcome instances?

The acquisition of these aspects is further utilized to describe the relationship between the dependent and independent variables (Kilic, 2013).

5.1.2 Random Forest

This classifier comprises a sequence of different tree classifiers where each classifier is developed by means of a random vector plotted independently from the input vector, and every single tree casts a single vote for the most widespread class to classify an input vector (Pal, 2005).

5.1.3 Decision Stump

A decision stump classifier is a kind of decision tree that employs a single attribute for splitting. In the case of discrete attributes, this simply implies that the tree only contains a single interior node, while the tree may be more complicated if the data attributes are numerical type (Fürnkranz, 2016).

5.1.4 Locally Weighted Learning

Locally weighted learning classification is a lazy approach, which continues the processing of trainset until the user interest is predicted. This processing typically involves storing the training data in the memory and finding relevant instances from the database to predict the user interest. As this classification technique utilizes memory for storing trainset thus, it is also called memory-based learning. It is a kind of nearest neighbor, and the relevance between two neighbors is mostly calculated through a distance function that focuses on neighboring points having high relevance (Atkeson et al., 1997).

5.2 Rating Prediction Metrics

Rating prediction metrics calculate the distance between predicted ratings and real ratings. The lower the value of errors, reveals highest the prediction accuracy. This work's predictive metrics include MAE, RMSE, RAE, RRSE, and elapsed time. All of these metrics measure the magnitude of errors in the predicted ratings. In contrast, the elapsed time refers to the total average time consumed by a classifier for training and testing processes.

5.3 Discussion of Results

The results were performed in Weka Software version 3.9. The experiments were performed in two phases; the first phase includes the experiments with all the classification algorithms compatible with the TripAdvisor dataset, while in the second phase top 5 classification algorithms were compared with each other. K-star, was selected as a primary approach to classify tourism attractions.

Table 2 and Table 3 provide the regression analysis results for the performance evaluation of all of these five algorithms. The comparatively k-star algorithm has the lowest error rate, which means it provides more accurate predictions for the TripAdvisor dataset.

Table 2 *Regression Analysis using Test/Train Percentage (80% - 20%)*

Algorithm	LR	RF	DS	LWL	K*
MAE	0.27	0.21	0.21	0.20	**0.17**
RMSE	0.32	0.25	0.25	0.24	**0.23**
RAE	99.73	76.87	77.48	72.59	**61.58**
RRSE	99.68	78.87	99.73	74.65	**70.68**
Elapsed Train Time	8.08	0.14	0.00	0.00	**0.00**
Elapsed Test Time	0.00	0.00	0.00	0.20	**0.93**

Table 2 reveals the results of the regression analysis of the k-star algorithm compared with other baseline approaches in terms of prediction evaluation metrics and total elapsed time.

In this regard, the data was split into two chunks, namely training set and testing set, where 784 (80%) instances out of 980 were considered training set while the remaining 196 (20%) instances were considered testing set. After splitting the dataset into train set and test set, the regression analysis was performed in Weka Experimenter. Moreover, each experiment was repeated five times and saved as comma-separated files. Each metric's average results were calculated using the excel formula for average and given in Table 2.

Table 2 illustrates that the k-star algorithm yields lower MAE, RMSE, RAE, and RRSE values, indicating fewer erroneous predictions for tourism attractions than other baseline approaches. In contrast, Linear Regression yields the highest values of these prediction metrics, indicating lower performance and higher error rates for the classification of tourism attractions using LR. Furthermore, k-star algorithms consume less time for training the model and high testing time which are good performance indicators, while Linear Regression performs vice versa.

Table 3 *Regression Analysis using 10-fold cross-validation*

Algorithm	LR	RF	DS	LWL	K*
MAE	0.27	0.20	0.21	0.19	**0.16**
RMSE	0.32	0.25	0.25	0.24	**0.22**
RAE	100.42	75.95	76.46	70.44	**58.19**
RRSE	100.46	76.47	78.13	73.66	**68.46**

Elapsed Train Time	10.77	0.26	0.00	0.00	**0.00**
Elapsed Test Time	0.00	0.00	0.00	0.16	**0.68**

After performing experiments on the train set and test set, the 10-fold cross-validation regression analysis was carried out to justify the performance of the k-star algorithm compared with other baseline algorithms. Furthermore, it can be observed from the results of table 3 that the second most efficient approach is LWL for the classification of tourism attractions using the TripAdvisor dataset.

6. Conclusion

Since the tourist attractions are getting the attention of people, researchers are motivated to discover the best approaches for the classification of tourism attractions and to predict/recommend the most suitable attraction on behalf of user interests. In this work, a classification technique called the k-star algorithm is applied for the prediction of tourism attraction based on user ratings. In this regard, a tourism dataset, TripAdvisor, is utilized for experimentations in this study. Different evaluation metrics were considered to justify the k-star algorithm's performance compared with other baseline approaches. The experimental evaluations of this work imply that the k-star algorithm outperformed other classification algorithms in terms of prediction evaluation metrics used in this study. K-star algorithm proved the best match for the TripAdvisor dataset classification, while in the case of other tourism datasets, it may provide different results.

References

[1]Assistant Professor, Department of Information Technology, University of the Punjab Gujranwala Campus, Gujranwala, 52250, Pakistan.

[2]Scholar, Department of Information Technology, University of the Punjab Gujranwala Campus, Gujranwala, 52250, Pakistan.

[3]Assistant Professor, Department of Information Technology, University of the Punjab Gujranwala Campus, Gujranwala, 52250, Pakistan.

[4]Lecturer, Department of Computer Science, Umm Al-Qura University, Makkah Al-Mukarmah, 24381, Saudi Arabia.

[5]Assistant Professor Smart Computing and Applied Sciences Group, Department of Physics, University of the Punjab, Lahore, 54590, Pakistan.

[6]Assistant Professor, Institute of Metallurgy & Materials Engineering, University of the Punjab, Lahore, 54590, Pakistan.

[*]Corresponding Author: Ghulam Mustafa. Email: gmustafa@pugc.edu.pk

Atkeson, C.G., Moore, A.W., Schaal, S., 1997. Locally Weighted Learning. Artif. Intell. Rev. 11, 11–73.

Bahri, A., Sugumaran, V., Devasenapati, S.B., 2013. Misfire Detection in IC Engine using Kstar Algorithm.

Cleary, J.G., Trigg, L.E., 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. Mach. Learn. Proc. 1995 108–114. https://doi.org/10.1016/b978-1-55860-377-6.50022-0

Fürnkranz, J., 2016. Decision Stump, in: Encyclopedia of Machine Learning and Data Mining. pp. 330–330. https://doi.org/10.1007/978-1-4899-7687-1_285

Jaynes, E.T., 1957. Information theory and statistical mechanics. II. Phys. Rev. 108, 171–190. https://doi.org/10.1103/PhysRev.108.171

Jegadeeshwaran, R., Sugumaran, V., 2014. Vibration Based Fault Diagnosis Study of an Automobile Brake System Using K Star (K*) Algorithm – A Statistical Approach. Recent Patents Signal Process. 4, 44–56. https://doi.org/10.2174/2210686304666140919011156

Kashevnik, A.M., Ponomarev, A. V., Smirnov, A. V., 2017. A multimodel context-aware tourism recommendation service: Approach and architecture. J. Comput. Syst. Sci. Int. 56, 245–258. https://doi.org/10.1134/S1064230717020125

Kavitha, S., Jobi, V., Rajeswari, S., 2017. Tourism recommendation using social media profiles. Adv. Intell. Syst. Comput. 517, 243–253. https://doi.org/10.1007/978-981-10-3174-8_22

Khallouki, H., Abatal, A., Bahaj, M., 2018. An ontology-based context awareness for smart tourism recommendation system. ACM Int. Conf. Proceeding Ser. https://doi.org/10.1145/3230905.3230935

Kilic, S., 2013. Linear regression analysis. J. Mood Disord. 3, 90. https://doi.org/10.5455/jmood.20130624120840

Kularbphettong, K., Ngamkam, B., 2014. A Recommendation System for Heritage-Tourism based on Mobile Application and Ontology Technique. Int. J. Inf. Process. Manag. 5, 42.

Lee, C.I., Hsia, T.C., Hsu, H.C., Lin, J.Y., 2017. Ontology-based tourism recommendation system. 2017 4th Int. Conf. Ind. Eng. Appl. ICIEA 2017 376–379. https://doi.org/10.1109/IEA.2017.7939242

Li, M., Vitanyi, P., 1997. An introduction to Kolmogorov complexity and its applications. Comput. Math. with Appl. 34, 137. https://doi.org/10.1016/s0898-1221(97)90213-3

Madhusudana, C.K., Kumar, H., Narendranath, S., 2016. Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal. Eng. Sci. Technol. an Int. J. 19, 1543–1551. https://doi.org/10.1016/j.jestch.2016.05.009

Painuli, S., Elangovan, M., Sugumaran, V., 2014. Tool condition monitoring using K-star algorithm. Expert Syst. Appl. 41, 2638–2643. https://doi.org/10.1016/j.eswa.2013.11.005

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222. https://doi.org/10.1080/01431160412331269698

Renjith, S., 2018. UCI Machine Learning Repository: Travel Reviews Data Set.

Renjith, S., Sreekumar, A., Jathavedan, M., 2018. Evaluation of partitioning clustering algorithms for processing social media data in tourism domain. 2018 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2018 127–131. https://doi.org/10.1109/RAICS.2018.8635080

Saputra, R.Y., Nugroho, L.E., Kusumawardani, S.S., 2019. Collecting the Tourism Contextual Information data to support the tourism recommendation system. 2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019 79–84. https://doi.org/10.1109/ICOIACT46704.2019.8938546

Satishkumar, R., Sugumaran, V., 2017. Remaining life time prediction of bearings using K-star algorithm – a statistical approach. J. Eng. Sci. Technol. 12, 168–181.

Smirnov, A., Kashevnik, A., Ponomarev, A., Shilov, N., Schekotov, M., Teslya, N., 2013. Recommendation system for tourist attraction information service. Conf. Open Innov. Assoc. Fruct 5, 148–155. https://doi.org/10.1109/FRUCT.2013.6737957

Tejera Hernández, D.C., 2015. An Experimental Study of K* Algorithm. Int. J. Inf. Eng. Electron. Bus. 7, 14–19. https://doi.org/10.5815/ijieeb.2015.02.03

UCI, 2018. UCI Machine Learning Repository [WWW Document]. URL https://archive.ics.uci.edu/ml/index.php (accessed 9.6.19).

Venkata Ramana, B., Babu, M.S.P., Venkateswarlu, N.., 2011. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. Int. J. Database Manag. Syst. 3, 101–114. https://doi.org/10.5121/ijdms.2011.3207

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Data Min. Pract. Mach. Learn. Tools Tech. 1–621.

Yee, C.N., Allison, L., 1993. Reconstruction of strings past. Bioinformatics 9, 1–7. https://doi.org/10.1093/bioinformatics/9.1.1